Kyiv School of Economics & Kyiv Economics Institute

DISCUSSION PAPER SERIES

A Bayesian Model of Sample Selection with a Discrete Outcome Variable: Detecting Depression in Older Adults

Maksym Obrizan

Kyiv School of Economics and Kyiv Economics Institute

DP# 41 **July 2011**

Kyiv School of Economics and Kyiv Economics Institute
Yakira St. 13, 3d floor, 04119 Kyiv, Ukraine
Phone: (+380 44) 492-8012, Fax: (+380 44) 492-8011
Internet: www.kse.org.ua, www.kei.org.ua
E-mail: info@kse.org.ua

A Bayesian Model of Sample Selection with a Discrete Outcome Variable: Detecting Depression in Older Adults*

Maksym Obrizan†

July 2, 2011

Abstract

Depression as a major mental illness among older adults has attracted a lot of research attention. However, the problem of sample selection, inevitable in most health surveys, has been largely ignored. To fill in this gap, this paper formally models selection into the sample jointly with a discrete outcome variable for depression. A Bayesian model of sample selection is developed from a multivariate probit by (i) allowing missing depression status for non-selected respondents, and (ii) using Cholesky factorization of the inverse variance matrix to avoid a Metropolis-Hastings step in the Gibbs sampler. Non-selected respondents are less likely to suffer from depression.

JEL classification numbers: C11, C35, I1

Keywords: Multivariate probit model; Sample selection; Bayesian methods; Gibbs sampler

*An earlier version of this paper benefited from insightful comments by seminar participants at the Department of Health Management and Policy at the University of Iowa (Iowa City, USA) in October of 2009 and Kyiv School of Economics/Kyiv Economics Institute (Kyiv, Ukraine) in March of 2010. John Geweke suggested the extension of the Chib and Greenberg (1998) paper, provided help on some derivations, and also offered financial support in the Spring of 2009. Micah Stack (Iowa City, USA) assisted in editing the manuscript. Any remaining errors are mine.

†Assistant Professor, Kyiv School of Economics (KSE) and Senior Economist, Kyiv Economics Institute (KEI), Kyiv, Ukraine, *Phone:* (+380) 44-492-8012, *Fax:* (+380) 44-492-8011, *Homepage:* http://mobrizan.weebly.com/, *E-mail:* mobrizan@kse.org.ua

1 INTRODUCTION

Mental illness constitutes a serious economic and health problem in developed countries. The disease burden of mental illness, measured as disability adjusted life years (DALYs), ranked second in the United States, with major depression being the leading cause within the mental health category (USDHHS, 1999). The prevalence of depression in older adults is of particular concern given the rapidly aging population. Based on estimates in the literature, McCall *et al.* (2002) report the prevalence rates of 2 to 6% for major depression and 8 to 20% for minor depression in older, community-based residents.

The prognosis of such depressive states is poor (Cole and Dendukuri, 2003). A meta-analysis of outcomes after 2 years estimated that only 33% of subjects were well, 33% were still depressed, and 21% had died (Cole *et al.*, 1999). In a longitudinal study of community-dwelling older adults, Cronin-Stubbs *et al.* (2000) found that even mild depression had an effect on changes in disability status, independent of physical health status and demographic factors. Furthermore, depression increases the perception of poor health (Wells and Burman, 1991), the utilization of medical services (Katon *et al.*, 1992), and health care costs (Unutzer *et al.*, 1997). Overall, depression constitutes a serious problem for older adults, leading to a reduced quality of life (Netuveli *et al.*, 2006). In light of these facts, it is not surprising that depression in older adults has been extensively studied, with risk factors being identified in Cole and Dendukuri (2003), among others.

Applied econometricians are well aware of the ubiquitous problem of sample selection in health surveys. For example, in the Survey on Assets and HEAlth Dynamics among the Oldest Old (AHEAD), 791 of the age-eligible participants (or 10.62%) are proxy respondents at the baseline. As such, they did not answer the affective health questions used to detect depression. This fact is not a flaw of the survey design because AHEAD - which is now a part of the Health and Retirement Study - is probably the best available longitudinal

health survey of community-dwelling older adults in the United States.[1]

But what if this 10% of the respondents are not selected at random? It looks quite plausible that the most depressed patients would not participate in the survey, for example, because they have lost interest in life. In other words, estimating the risk factors for depression by simply excluding non-responding patients may lead to biased and inconsistent estimates. Similar arguments have been widely used by applied econometricians who adopted Heckman's (1979) sample selection model.

The novelty of this work, however, is that relatively few applications have been considered for the case of a *binary* dependent variable. Following the medical literature (Covinsky *et al.*, 2006; Choi and Kim, 2007), depression in this study is defined as having 3 or more symptoms on a shortened 8-item version of the CES-D (Center for Epidemiological Studies Depression Scale), which was used in all waves of the AHEAD study (Sheffick, 2000). Unlike Heckman's (1979) original formulation, the dependent variable takes only two values: "depressed" for 1,886 respondents and "not depressed" for 3,759 participants. The remaining 1,802 respondents have a depression status of "unobserved".

The earliest application of Heckman's (1979) model to a discrete outcome variable is Wynand and Praag (1981). It was further used in Meng and Schmidt (1985) and Mohanty (2002) provided a recent application to teen employment. Another relevant example in classical econometrics is Greene (1992), who refers to an earlier paper by Boyes, Hoffman and Low (1989). Kenkel and Terza (2001) use a two-step estimator in the model of alcohol consumption (number of drinks) with an endogenous dummy for advice (from a physician to reduce alcohol consumption). Munkin and Trivedi (2003) discuss the problems of different classical estimators of selection models using the discrete outcome equation.[2]

[1]The HRS is sponsored by the National Institute of Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan.

[2]The journal format does not allow for an in-depth discussion of the previous literature, which is available from the author's earlier MPRA Paper No. 28577.

The Bayesian model of sample selection developed herein instead relies on *data augmentation*, *Gibbs sampling* and joint estimation of the selection and outcome equations, as do other recent Bayesian treatments (Li, 1998; Huang, 2001; van Hasselt, 2008). Latent variables corresponding to observed binary outcomes are treated as additional parameters and are sampled from the joint posterior distribution.

The starting point of the study is the multivariate probit model of Chib and Greenberg (1998), albeit with two extensions. The first extension allows for missing outcomes when depression status is "unobserved". In terms of latent variable representation, this means that multivariate normal distribution is not truncated in the direction of a missing outcome. The second extension is a convenient Cholesky factorization of the inverse variance matrix that allows for avoidance of the Metropolis-Hastings step in the Gibbs sampler.[3]

The joint estimation of a binary selection equation ("selected", "not selected") and a discrete outcome equation ("depressed", "not depressed", "unobserved") reveals that the correlation coefficient between the error terms is not zero. In addition, an extended multivariate probit model demonstrates proper convergence, which makes it applicable in other empirical problems with a discrete outcome that are subject to sample selection.

The rest of the paper is organized as follows. Section 2 discusses two extensions of the multivariate probit model and sets up the Gibbs sampler. Section 3 describes the data and provides the summary statistics. Section 4 provides the details of the estimation results, and the last section concludes.

[3]Alternative extensions to Chib and Greenberg (1998) multivariate probit model appear in Chib and Hamilton (2000, 2002), where a discrete treatment variable selects one of the two potential outcomes. In these models, at least one potential outcome is observed, contrary to the current setup where a depression status is missing for "not selected" respondents.

2 THE ECONOMETRIC MODEL

2.1 Two extensions to the multivariate probit model

Chib and Greenberg (1998) offer the following Bayesian treatment of the multivariate probit model. Suppose that a researcher observes a set of potentially correlated binary events $i = 1, \ldots, m$ over an independent sample of $t = 1, \ldots, T$ respondents. For each respondent, t, there exists a corresponding vector of latent variables $\widetilde{y}_{.t} = (\widetilde{y}_{1t}, ..., \widetilde{y}_{mt})'$ that is assumed to follow a normal distribution

$$\widetilde{y}_{.t} \sim N_m(Z_t\beta, \Sigma), \tag{1}$$

where $Z_t = diag(Z_{1t}, ..., Z_{mt})$ is an $m \times k$ covariate matrix, $\beta_i \, \epsilon \, R^{k_i}$ is an unknown parameter vector in equation $i = 1, ..., m$ with $\beta = (\beta_1', ..., \beta_m')' \, \epsilon \, R^k$ and $k = \sum_{i=1}^{m} k_i$, and Σ is the variance matrix. There is potential correlation in the disturbance terms for respondent t across events $i = 1, ..., m$ coming from some unobserved factor that simultaneously affects multiple dependent variables.

The sign of $\widetilde{y}_{it}$ for each dependent variable $i = 1, ..., m$ uniquely determines the observed binary outcome y_{it}:

$$y_{it} = I(\widetilde{y}_{it} > 0) - I(\widetilde{y}_{it} <= 0) \quad (i = 1, ..., m), \tag{2}$$

where $I(A)$ is the indicator function of an event A. The probability of observing a vector of binary responses $Y_{.} = (Y_1, ..., Y_m)'$ for individual t can be expressed as

$$\int_{B_{mt}} ... \int_{B_{1t}} \phi_m(\widetilde{y}_{.t}|Z_t\beta, \Sigma) d\widetilde{y}_{.t}, \tag{3}$$

where $B_{it} \, \epsilon \, (0, \infty)$ if $y_{it} = 1$ and $B_{it} \, \epsilon \, (-\infty, 0]$ if $y_{it} = -1$. To avoid difficulties associated with evaluating multivariate normal integral by conventional methods, Chib and Greenberg (1998) refer to a Bayesian technique of *data augmentation* to simulate the latent variable $\widetilde{y}_{.t}$ from the conditional posterior distribution.

Unfortunately, the formulation in Chib and Greenberg (1998) cannot be directly used to

estimate the model of interest in this paper. In the current setup, a binary selection variable ("selected", "not selected") is observed for all respondents t.[4] However, a discrete outcome variable ("depressed", "not depressed", "unobserved") is always "unobserved" for "not selected" respondents. To overcome this difficulty, the unobserved outcome variable $\widetilde{y}_{it}$ is not restricted and can take any value in the interval $(-\infty, \infty)$.

The second extension to the setup in Chib and Greenberg (1998) allows the researcher to avoid the Metropolis-Hastings algorithm for drawing the elements of the variance matrix. In particular, this paper uses the Cholesky factorization of the inverse of the variance matrix $\Sigma^{-1} = \breve{F} \cdot \breve{F}'$ where $\breve{F}$ is the lower triangular matrix. If the diagonal elements of $\breve{F}$ are arrayed in a diagonal matrix Q, then $\Sigma^{-1} = \breve{F} Q^{-1} Q^2 Q^{-1} \breve{F}' = F Q^2 F$. Then the variance matrix is defined by F, which is a lower triangular matrix that has ones on the main diagonal,

$$F = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ f_{21} & 1 & 0 & \cdots & 0 \\ f_{31} & f_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{m1} & f_{m2} & f_{m3} & \cdots & 1 \end{pmatrix}.$$

and $D^{-1} = I(m)$ which is an identity matrix. This latter condition is needed for the identification of the multivariate probit model that has only $m(m-1)/2$ identified parameters in the variance matrix.

2.2 The Gibbs Sampler

Given a prior density $p(\beta, F, D)$ on the parameters β, F and D, the posterior density is equal to

$$p(\beta, F, D|y) \propto p(\beta, F, D) p(y|\beta, \Sigma), \tag{4}$$

[4] As well as all independent variables in both equations.

where $p(y|\beta,\Sigma)=\prod_{t=1}^{T}p(y_{.t}|\beta,\Sigma)$ is the likelihood function.

In this representation, the evaluation of the likelihood function by conventional methods may be computationally intensive. Albert and Chib (1993) developed an alternative Bayesian framework that focuses on the joint posterior distribution of the parameters and the latent data $p(\beta,F,D,\widetilde{y}_1,...,\widetilde{y}_T|y)$ such that

$$\begin{aligned} p(\beta,F,D,\widetilde{y}|y) &\propto p(\beta,F,D)p(\widetilde{y}|\beta,\Sigma)p(y|\widetilde{y},\beta,\Sigma) \\ &= p(\beta,F,D)p(\widetilde{y}|\beta,\Sigma)p(y|\widetilde{y}). \end{aligned} \tag{5}$$

It is possible now to implement a sampling approach and construct a Markov chain from the distributions $[\widetilde{y}_{.t}|y_{.t},\beta,\Sigma]$ $(t\leq T)$, $[\beta|y,\widetilde{y},\Sigma]$ and $[F,D|y,\widetilde{y},\beta]$.

The parameters in β and F are specified to be independent in the prior. The prior distribution for β is assumed to be normal $\phi_k(\beta|\underline{\beta},\underline{B}^{-1})$, with the location vector $\underline{\beta}$ and the precision matrix $\underline{B}$.

It is convenient to concatenate the vectors below the main diagonal in the F matrix as

$$F_{vector}=\begin{pmatrix} F_{2:m,1} \\ F_{3:m,2} \\ \vdots \\ F_{m,m-1} \end{pmatrix},$$

where $F_{i+1:m,i}$ for $i=1,...,m-1$ represents elements from $i+1$ to m in column i. The prior distribution of F_{vector} is assumed to be $\left(\frac{m(m-1)}{2}\right)$-variate normal

$$F_{vector}\sim N(\underline{F}_{vector},\underline{H}^{-1}). \tag{6}$$

In this expression $\underline{F}_{vector}$ is the prior mean of the normal distribution, and the prior variance matrix $\underline{H}^{-1}$ is block-diagonal with

$$\underline{H}=\begin{pmatrix} \underline{H}_{2:m,2:m} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \underline{H}_{3:m,3:m} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \underline{H}_{1,1} \end{pmatrix}.$$

This precision matrix has $(m-1)\times(m-1)$ matrix $\underline{H}_{2:m,2:m}$ in the upper left corner, and the matrix dimension decreases by one in each consequent block on the main diagonal. The lower right matrix $\underline{H}_{1,1}$ is a scalar.

A Gibbs sampler is constructed by drawing from the following conditional posterior distributions: the vector of coefficients β, the F_{vector} from the variance matrix decomposition and the latent vector $\widetilde{y}_{.t}$ for each respondent $t \leq T$.[5]

The posterior distribution of β comes from the posterior density kernel and is normal

$$\beta|(\widetilde{y},\Sigma) \sim N_k(\beta|\overline{\beta},\overline{B}^{-1}), \tag{7}$$

where $\overline{B} = \underline{B} + \sum_{t=1}^{T} Z_t'\Sigma^{-1}Z_t$ and $\overline{\beta} = \overline{B}^{-1}(\underline{B}\underline{\beta} + \sum_{t=1}^{T} Z_t'\Sigma^{-1}\widetilde{y}_{.t})$.

The posterior conditional distribution of F_{vector} is also normal

$$F_{vector}|(y,\widetilde{y},\beta) \sim N_{\left(\frac{m(m-1)}{2}\right)}(\overline{F}_{vector},\overline{H}^{-1}). \tag{8}$$

The conditional posterior normal distribution has the posterior precision matrix

$$\overline{H} = \underline{H} + \begin{pmatrix} \sum_{t=1}^{T}\varepsilon_{t,2:m}\varepsilon_{t,2:m}' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sum_{t=1}^{T}\varepsilon_{t,3:m}\varepsilon_{t,3:m}' & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \sum_{t=1}^{T}\varepsilon_{t,m}\varepsilon_{t,m}' \end{pmatrix}.$$

The posterior mean of the normal distribution is equal to

$$\overline{F}_{vector} = \overline{H}^{-1}\underline{H}\ \underline{F}_{vector} - \overline{H}^{-1}\begin{pmatrix} \sum_{t=1}^{T}\varepsilon_{t,2:m}\varepsilon_{t,1} \\ \sum_{t=1}^{T}\varepsilon_{t,3:m}\varepsilon_{t,2} \\ \vdots \\ \sum_{t=1}^{T}\varepsilon_{t,m}\varepsilon_{t,m-1} \end{pmatrix}.$$

Finally, the latent data $\widetilde{y}_{.t}$ are drawn independently for each respondent $t \leq T$ from the truncated multivariate normal distribution as described in Geweke (1991). The algorithm

[5]Complete details of the Gibbs sampler derivation are available from the Appendix.

makes draws conditional on Z_t, β and F as well as $\widetilde{y}_{.t}$ obtained in the previous draw. The multivariate normal distribution is truncated to the region defined by the $m \times 2$ matrix $[a, b]$ with a typical row i equal to $(0, \infty)$, if $y_{it} = 1$ and $(-\infty, 0)$ if $y_{it} = -1$. If y_{it} is not observed, then row i is $(-\infty, \infty)$.

3 THE DATA

This paper employs data from the Survey on Assets and HEAlth Dynamics among the Oldest Old (AHEAD), which is a nationally representative longitudinal health and retirement study administered by the Survey Research Center at the University of Michigan. A complete description of the AHEAD study can be found online at http://hrsonline.isr.umich.edu and in several review articles (Juster and Suzman, 1995; Myers *et al.*, 1997).

Depression is measured using a shortened 8-item version of the CES-D (Center for Epidemiological Studies Depression Scale), which was used in all waves of the AHEAD study (Steffick, 2000). Prior research indicates that a shorter version of the original 20-item CES-D is appropriate for assessing depressive symptoms in older population, and it can also alleviate respondent burden (Kohout *et al.*, 1993; Fonda and Herzog, 2001). In the previous studies, the cut-off of 3 or more on the 8-item CES-D scale was used as a threshold of depression (Steffick, 2000; Choi and Kim, 2007; Covinsky *et al.*, 2006). Thus, depression was defined as having three or more CES-D items at the last available follow-up interview after the baseline. Out of 5,645 respondents with a CES-D count, 3,759 (66.4%) were "not depressed" and 1,886 (33.6%) were "depressed", according to the adopted definition. There were also 1,802 respondents with a depression status of "unobserved", which constitutes 24.2% of the original AHEAD sample of 7,447 age-eligible older adults. The primary reasons for having an "unobserved" depression status are (i) being a proxy respondent at the baseline interview and (ii) participating only in the baseline survey but not in any of the follow-ups. Thus, selection variable takes value of "observed" for 5,645 respondents and "unobserved" for the remaining

1,802 participants.

The list of covariates for the depression equation includes available predictors found in previous studies using the CES-D scale as identified in a review article by Cole and Dendukuri (2003). Age and number of years in the sample are normalized by the corresponding means. Age varies from 69 to 103 years at the baseline interviews in 1993-1994, with a mean of 77 years. Participants stayed in the sample from 2 to 11 years with a mean of 6 years. The list of covariates for the selection equation is partially based on the author's prior experience with modeling sample selection in the AHEAD sample (Kaskie BP *et al.*, 2010). All missing observations for independent variables are set to the corresponding medians. Table 1 provides complete details on covariates.

4 RESULTS

The implementation of the Gibbs sampler is programmed in the Matlab environment with some loops written in C language. All the codes successfully passed the joint distribution tests in Geweke (2004). The prior distribution of F_{vector} is assumed to be normal, with a mean of zero and variance of 10. The location vector $\underline{\beta}$ is set to zeros and the precision matrix $\underline{B}$ is set using the g-prior of Zellner (1986), which allows for the proper scaling of parameters relative to each other. In particular, the variance matrix $\underline{B}^{-1}$ is block diagonal, with each block $i = 1, 2$ defined as $g_B \cdot (Z_i' Z_i)^{-1}$ of the corresponding dimension $k_i \times k_i$. In this expression, Z_i is a $T \times k_i$ set of explanatory variables in equation i. The scaling factor g_B is set to $T \cdot m/k$ where $T = 7,447$ observations, $m = 2$ equations, and $k = 30$ variables. Extensive sensitivity analysis indicates that results are not driven by the choice of the prior. Both selection and outcome equations contain intercept and 14 binary, discrete and continuous covariates. The results are based on 20,000 draws from the Gibbs sampler, with the initial 20% dropped as burn-in iterations. To reduce the problem of autocorrelation, only every 10th draw is retained, resulting in 1,600 draws used in computations of posterior

Table 1: Descriptive statistics for 7,447 respondents in the sample

Variable	Mean	Std	# Missing
Age (norm)	0.000	5.908	0
Age squared (norm)	0.000	946.840	0
Years in sample (norm)	0.000	4.229	618
Years in squared (norm)	0.000	50.291	618
Men	0.391	0.488	2
African American	0.136	0.343	0
Hispanic	0.056	0.230	0
Lives alone (1993)	0.358	0.479	0
Widowed (1993)	0.408	0.491	2
Completed grade school	0.281	0.449	0
Completed high school	0.291	0.454	0
Household Income (1993)	24839	50054	2
Central region of US (1993)	0.258	0.437	0
East region of US (1993)	0.196	0.397	0
West region of US (1993)	0.156	0.363	0
Body-mass index (1993)	25.362	4.503	79
Diabetes ever	0.179	0.384	628
Psychological problems ever	0.109	0.311	0
Stroke ever	0.169	0.375	626
EVGG health (1993)	0.630	0.483	0
EVGG vision (1993)	0.726	0.446	0
# of ADL difficulties (1993)	0.389	0.932	3
# of IADL difficulties (1993)	0.493	1.117	3

Note: "Norm" - normalized at means,"EVGG" - Excellent, Very Good or Good.

moments. The coefficients in the first (selection) equation are normalized by $\sqrt{1+F^2(j)}$ in each draw $j = 1, \ldots, 20,000$ to have variances comparable with the second equation.

Table 2 reports selected posterior moments for coefficients in both equations. Geweke's

(1992) convergence diagnostic test does not indicate any convergence problems in any of the coefficients in the two equations. Figure 1 depicts the behavior of 5 selected coefficients from the depression equation. It appears that the multivariate probit model with sample selection works well in terms of its convergence properties.

Table 2. The posterior statistics for β coefficients.

Variable	pmean	pstd	CD	pmean	pstd	CD
Intercept	-0.070	0.105	1.393	-0.509	0.069	0.146
Age (norm)	0.083	0.058	-0.044	0.107	0.081	0.019
Age squared (norm)	-0.001	0.000	-0.081	-0.001	0.001	-0.024
Years in sample (norm)				0.514	0.085	-0.022
Years in squared (norm)				-0.035	0.006	0.062
Men	-0.150	0.035	-0.648	-0.140	0.040	-0.198
African American	-0.198	0.050	0.888	0.110	0.052	0.643
Hispanic	-0.125	0.074	-0.860	0.199	0.079	-1.139
Lives alone (1993)	0.187	0.036	-0.631			
Widowed (1993)				-0.044	0.041	-0.092
Completed grade school	-0.305	0.042	0.718			
Completed high school	-0.054	0.042	0.601			
Household Income (1993)	0.000	0.000	0.328			
Central region of US (1993)	0.101	0.044	-0.646			
East region of US (1993)	-0.024	0.046	-0.585			
West region of US (1993)	0.036	0.050	0.418			
Body-mass index (1993)	0.029	0.004	-1.299			
Diabetes ever				0.046	0.047	-1.151
Psychological problems ever				0.372	0.055	0.082
Stroke ever				0.226	0.049	-0.746
EVGG health (1993)				-0.450	0.040	0.456
EVGG vision (1993)	0.242	0.038	-0.016			
# of ADL difficulties (1993)				0.144	0.027	-0.605
# of IADL difficulties (1993)				-0.092	0.043	0.511
ρ				0.163	0.035	0.116

Note: "pmean" and "pstd" stand for the posterior mean and standard deviation, "CD" stands for Geweke's (1992) convergence diagnostic statistics. Columns 2-4 refer to the selection equation and columns 5-7 refer to the depression equation.

Simple manipulation with the covariance matrix shows that the correlation coefficient be-

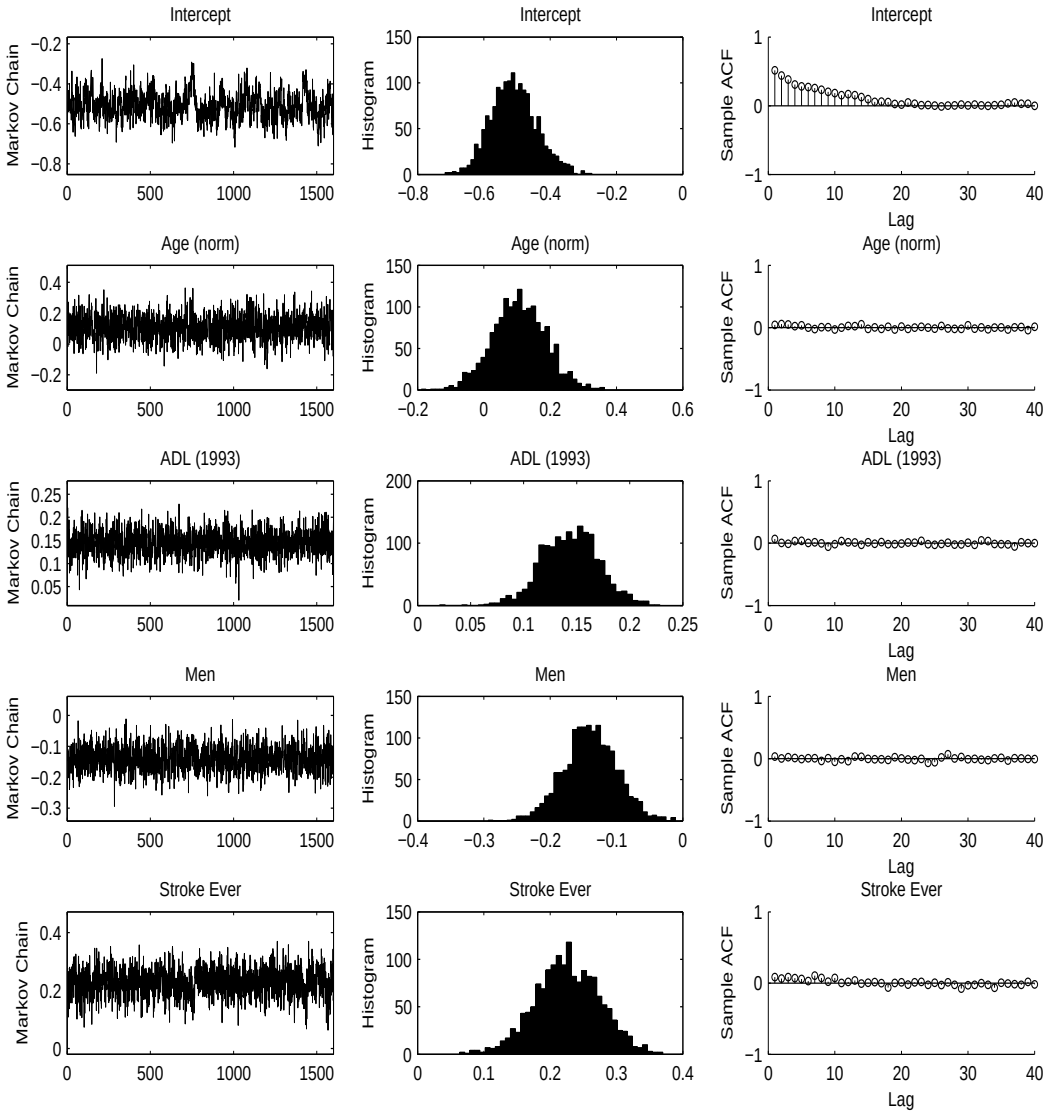

Figure 1: Selected coefficients from the depression equation.

tween the error terms in two equations can be computed as

$$\rho(j) = \frac{-F(j)}{\sqrt{1 + F(j)^2}}, \tag{9}$$

for each iteration $j = 1, \ldots, 20,000$. The correlation coefficient is statistically different from zero judging from a 99% highest density posterior interval (HDPI). This indicates the presence of unobserved effect and, hence, sample selection bias in the depression equation. Although the value of the correlation coefficient is only 0.16, sample selection may bias the coefficient estimates. Positive correlation implies, somewhat surprisingly, that the unobserved factor increases the probability of being selected into the sample and having depression.

The posterior moments from the selection equation indicate that older people, men, African Americans, and respondents who completed grade school only are less likely to be selected for the sample. On the other hand, respondents who live alone, have a higher body-mass index, have a higher household income, live in the Central region of the US, and have good to excellent vision are more likely to be included in the sample. The importance of socioeconomic factors for sample selection is consistent with author's prior experience using the AHEAD sample (Kaskie *et al.*, 2010).

Unlike previous studies of depression, the current paper considers the effect of sample selection formally, implying that the results should be subject to lower bias. The strongest predictor of depression is number of years participating in the survey, which is subject to diminishing effect as can be judged from the squared term. The self-reported (true or perceived) good through excellent health has the strongest protective effect. Women, African Americans, and Hispanics are at elevated risk for depression. Stroke and prior psychological problems can also increase the probability of developing depression at the last available interview after the baseline. Each additional limitation in the 5 Activities of Daily Living (getting across a room, dressing, bathing or showering, eating, and getting into or out of bed) increases the probability of depression. On the other hand, additional limitations in the 5 Instrumental Activities of Daily Living (using a telephone, taking medication, handling

money, shopping, and preparing meals) reduce the chance of developing depression.

5 CONCLUSIONS

This paper is the first attempt to formally account for the sample selection bias in a study of depression in older, community-dwelling Americans. Although the selection equation contains a large set of predictors for being included in the sample, the unobserved effect remains significant, as indicated by the positive correlation coefficient. Thus, excluded participants are less likely to suffer from depression, leading to possible overestimation of the depression rates among the elderly in the reported studies.

Women, minorities, and longer participation in the survey are primary sociodemographic predictors for developing depression. Prior stroke and psychological difficulties, as well as higher counts of limitations in Activities of Daily Living, are health factors associated with higher probability of depression.

In addition, this is the first application of a novel sample selection model based on the Bayesian multivariate probit setup utilized by Chib and Greenberg (1998). The model demonstrates proper convergence and stability. Although the model in this paper was specifically developed for a health application, it can be readily used whenever sample selection might be an issue and outcome variable is discrete.

APPENDIX A: DERIVATIONS OF THE CONDITIONAL POSTERIOR DISTRIBUTIONS

This Appendix derives the conditional posterior distributions in the Gibbs sampler. The posterior density kernel is the product of the prior for β, prior for F_{vector} and augmented

likelihood for $\widetilde{y}'_{.t}s$

$$
\begin{aligned}
& |\underline{B}|^{1/2} \exp\Big\{ -\frac{1}{2}(\beta-\underline{\beta})'\underline{B}(\beta-\underline{\beta})\Big\} \qquad (10)\\
\cdot\ & |\underline{H}|^{1/2} \exp\Big\{ -\frac{1}{2}(F_{vector}-\underline{F}_{vector})'\underline{H}(F_{vector}-\underline{F}_{vector})\Big\}\\
\cdot\ & |\Sigma|^{-T/2}\prod_{t=1}^{T}\exp\Big\{ -\frac{1}{2}(\widetilde{y}_{.t}-Z_t\beta)'\Sigma^{-1}(\widetilde{y}_{.t}-Z_t\beta)\Big\}I(\widetilde{y}_{.t}\ \epsilon\ B_t).
\end{aligned}
$$

The three conditional posterior distributions can be obtained as follows.

(i) The conditional posterior kernel for β can be obtained from equation (10) by collecting the terms that contain β and completing the square

$$
\begin{aligned}
p(\beta|\Sigma,\widetilde{y}) \ &\propto\ \exp\Big\{ -\frac{1}{2}(\beta'\underline{B}\beta-2\beta'\underline{B}\underline{\beta}+\underline{\beta}'\underline{B}\underline{\beta})\Big\} \qquad (11)\\
&\cdot\ \prod_{t=1}^{T}\exp\Big\{ -\frac{1}{2}(\widetilde{y}_{.t}\Sigma^{-1}\widetilde{y}_{.t}-2\beta'Z_t'\Sigma^{-1}\widetilde{y}_{.t}+\beta'Z_t'\Sigma^{-1}Z_t\beta)\Big\}\\
&\propto\ \exp\Big\{ -\frac{1}{2}\Big(\beta'(\underline{B}+\sum_{t=1}^{T}Z_t'\Sigma^{-1}Z_t)\beta-2\beta'(\underline{B}\underline{\beta}+\sum_{t=1}^{T}Z_t'\Sigma^{-1}\widetilde{y}_{.t})\Big)\Big\}\\
&\propto\ \exp\Big\{ -\frac{1}{2}(\beta-\overline{\beta})'\overline{B}(\beta-\overline{\beta})\Big\},
\end{aligned}
$$

where $\overline{B}=\underline{B}+\sum_{t=1}^{T}Z_t'\Sigma^{-1}Z_t$ is the posterior precision and

$$\overline{\beta}=\overline{B}^{-1}(\underline{B}\underline{\beta}+\sum_{t=1}^{T}Z_t'\Sigma^{-1}\widetilde{y}_{.t})$$

is the posterior mean for β.

(ii) The alternative expression for the density of $\widetilde{y}$ is given by

$$
p(\widetilde{y}|y,\beta,F,D)\ \propto\ \prod_{i=1}^{m}\exp\Big\{ -\frac{1}{2}\sum_{t=1}^{T}(\varepsilon_{t,i}+F'_{i+1:m,i}\varepsilon_{t,i+1:m})^2\Big\}. \qquad (12)
$$

Remembering that $F_{vector} = [F'_{2:m,1}, ..., F'_{m,m-1}]'$ one can collect the terms in the posterior density kernel (10) as

$$
\begin{aligned}
p(F_{vector}|\beta,\widetilde{y}) &\propto \prod_{i=1}^{m-1} \exp\Big\{ -\frac{1}{2}(F_{i+1:m,i} - \underline{F}_{i+1:m,i})'\underline{H}_i(F_{i+1:m,i} - \underline{F}_{i+1:m,i})\Big\} \qquad (13)\\
\cdot \quad & \prod_{i=1}^{m-1} \exp\Big\{ -\frac{1}{2}\sum_{t=1}^{T}(\varepsilon_{t,i} + F'_{i+1:m,i}\varepsilon_{t,i+1:m})^2\Big\}\\
\propto \quad & \prod_{i=1}^{m-1} \exp\Big\{ -\frac{1}{2}\Big(F'_{i+1:m,i}\underline{H}_i F_{i+1:m,i} - 2F'_{i+1:m,i}\underline{H}_i\underline{F}_{i+1:m,i} + \underline{F}_{i+1:m,i}\underline{H}_i\underline{F}_{i+1:m,i}\Big)\\
\cdot \quad & \prod_{i=1}^{m-1} \exp\Big\{ -\frac{1}{2}\Big(\sum_{t=1}^{T}\varepsilon_{t,i}^2 + 2F'_{i+1:m,i}\sum_{t=1}^{T}\varepsilon_{t,i+1:m}\varepsilon_{t,i}\\
& + F'_{i+1:m,i}\big(\sum_{t=1}^{T}\varepsilon_{t,i+1:m}\varepsilon'_{t,i+1:m}\big)F_{i+1:m,i}\Big)\Big\}\\
\propto \quad & \prod_{i=1}^{m-1} \exp\Big\{ -\frac{1}{2}\Big(F'_{i+1:m,i}\big(\underline{H}_i + \sum_{t=1}^{T}\varepsilon_{t,i+1:m}\varepsilon'_{t,i+1:m}\big)F_{i+1:m,i}\\
& - 2F'_{i+1:m,i}\big(\underline{H}_i\underline{F}_{i+1:m,i} - \sum_{t=1}^{T}\varepsilon_{t,i+1:m}\varepsilon_{t,i}\big)\Big)\Big\}\\
\propto \quad & \prod_{i=1}^{m-1} \exp\Big\{ -\frac{1}{2}\Big(F_{i+1:m,i} - \overline{F}_{i+1:m,i})'\overline{H}_i(F_{i+1:m,i} - \overline{F}_{i+1:m,i}\Big)\Big\}
\end{aligned}
$$

where $\overline{H}_i = \underline{H}_i + \sum_{t=1}^{T}\varepsilon_{t,i+1:m}\varepsilon'_{t,i+1:m}$ is the posterior precision and

$$\overline{F}_{i+1:m,i} = \overline{H}_i^{-1}\underline{H}_i\ \underline{F}_{i+1:m,i} - \overline{H}_i^{-1}\sum_{t=1}^{T}\varepsilon_{t,i+1:m}\varepsilon_{t,i}$$

is the posterior mean. It is understood that $\underline{H}_i$ is the ith element of the block-diagonal prior precision matrix $\underline{H}$ with dimensions decreasing from $(m-1)\times(m-1)$ for the first block to 1×1 for the last block. The final step is to organize $i = 1 : m-1$ multivariate normal distributions in the last line of equation (13) into one

$$
\begin{aligned}
p(F_{vector}|\beta,\widetilde{y}) &\propto \prod_{i=1}^{m-1} \exp\Big\{ -\frac{1}{2}\Big(F_{i+1:m,i} - \overline{F}_{i+1:m,i})'\overline{H}_i(F_{i+1:m,i} - \overline{F}_{i+1:m,i}\Big)\Big\}\\
&= \exp\Big\{ -\frac{1}{2}\Big(F_{vector} - \overline{F}_{vector})'\overline{H}(F_{vector} - \overline{F}_{vector}\Big)\Big\}, \qquad (14)
\end{aligned}
$$

which is used in the text. Since $F_{vector} = (F'_{2:m,1}, F'_{3:m,2}, ..., F'_{m,m-1})'$ with $F'_{j+1:m,j} = (f_{j+1,j}, ..., f_{m,j})$ for $j = 1, ..., m-1$ being the vectors under the main diagonal of F

$$F = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ f_{21} & 1 & 0 & \cdots & 0 \\ f_{31} & f_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{m1} & f_{m2} & f_{m3} & \cdots & 1 \end{pmatrix}.$$

one can construct the covariance matrix Σ as

$$\Sigma = (F')^{-1}F^{-1}. \tag{15}$$

(iii) Finally, the latent data $\widetilde{y}_{.t}$ are drawn for each respondent $t \leq T$ from the truncated multivariate normal distribution as in Geweke (1991) conditional on Z_t, β and F, as well as $\widetilde{y}_{.t}$ obtained in the previous draw. The multivariate normal distribution is truncated to the region defined by the $m \times 2$ matrix $[a, b]$ with a typical row i equal to $(0, \infty)$ if $y_{it} = 1$ and $(-\infty, 0)$ if $y_{it} = -1$. If y_{it} is not observed, then row i is $(-\infty, \infty)$

REFERENCES

Boyes W, Hoffman D, Low S. 1989. An econometric analysis of the bank credit scoring problem. *Journal of Econometrics* 40:3-14.

Cole MG, Bellavance F, Mansour A. 1999. Prognosis of depression in elderly community and primary care populations: a systematic review and meta-analysis. *Am J Psychiatry* 156:1182-1189.

Cole MG, Dendukuri N. 2003. Risk factors for depression among elderly community subjects: A systematic review and meta-analysis. *American Journal of Psychiatry* 50, 1147-1156.

Covinsky KE *et al.* 2006. Development and validation of an index to predict Activity of Daily Living dependence in community-dwelling elders. *Med Care* (44):149-157.

Chib S, Greenberg E. 1998. Analysis of multivariate probit models. *Biometrika* 85(2):347-361.

Chib S, Hamilton BH. 2000. Bayesian analysis of cross-section and clustered data treatment models. *Journal of Econometrics* 97: 25-50.

Chib S, Hamilton BH. 2002. Semiparametric Bayes analysis of longitudinal data treatment models. *Journal of Econometrics* 110:67-89.

Choi NG, Kim FS. 2007. Age group differences in depressive symptoms among older adults with functional impairments. *Health & Social Work* 3(3): 177-188.

Fonda SJ, Herzog AR. 2001. Patterns and risk factors of change in somatic and mood symptoms among older adults. *Annals of Epidemiology* 11, 361-368.

Geweke J. 1991. Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints. In: E. Keramidas (ed.) Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface. Fairfax: Interface Foundation of North America, Inc.

Geweke J. 1992. Evaluating the accuracy of sampling-based approaches to the calculation of the posterior moments. In: J. Berger, J. Dawid and A. Smith (eds.) Bayesian Statistics 4. Oxford: Clarendon Press.

Geweke J. 2004. Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association* 99:799-804.

Greene W. 1992. A statistical model for credit scoring. WP No. EC-92-29, Department of Economics, Stern School of Business, New York University.

Juster FT, Suzman RM. 1995. An overview of the Health and Retirement Study. *J Hum Resources* 30: S7-S56.

Heckman J. 1979. Sample selection bias as a specification error. *Econometrica* 47(1):153-161.

Huang HC. 2001. Bayesian analysis of the SUR tobit model. *Applied Economics Letters* 8:617-622.

Kaskie *et al.* 2010. Defining Emergency Department Episodes by Severity and Intensity: A 15-year Prospective Study of Medicare Beneficiaries. *BMC Health Services Research* 10:173.

Katon *et al.* 1992. Adequacy and duration of antidepressant treatment in primary care. *Med Care* 30:67Ű76.

Kenkel D, Terza J. 2001. The effects of physician advice on alcohol consumption: Count regression with an endogenous treatment effect. *Journal of Applied Econometrics* 16(2):165-184.

Kohout F *et al.* 1993. Two shorter forms of the CES-D depression symptoms index. *Journal of Aging and Health* 5, 179-193.

Li K. 1998. Bayesian inference in a simultaneous equation model with limited dependent variables. *Journal of Econometrics* 85:387-400.

McCall *et al.* 2002. The prevalence of major depression or dysthymia among aged Medicare Fee-for-Service beneficiaries. *Int J Geriatr Psychiatry* 17: 557Ű565.

Myers GC, Juster FT, Suzman RM. 1997. Assets and Health Dynamics among the Oldest Old (AHEAD): Initial results from the longitudinal study. *J Gerontol Psychol Sci Soc Sci* 52B: Special Issue.

Meng C, Schmidt P. 1985. On the cost of partial observability in the bivariate probit model. *International Economic Review* 26:71-86.

Mohanty M. 2002. A bivariate probit approach to the determination of employment: a study of teen employment differentials in Los Angeles county. *Applied Economics* 34(2):143-156.

Munkin M, Trivedi P. 2003. Bayesian analysis of a self-selection model with multiple outcomes using simulation-based estimation: An application to the demand for healthcare. *Journal of Econometrics* 114:197-220.

Netuveli G *et al.* 2006. Quality of life at older ages: evidence from the English longitudinal study of aging (wave 1). *J Epidemiol Community Health* 60:357-363.

Steffick DE. 2000. HRS/AHEAD documentation report: Documentation of affective functioning measures in the Health and Retirement Study. Ann Arbor: University of Michigan, Survey Research Center.

Unutzer J *et al.* 1997. Depressive symptoms and the cost of health services in HMO patients aged 65 and over: a 4-year prospective study. *JAMA* 277:1618-1623.

US Department of Health and Human Services. 1999. *Mental Health: A Report of the Surgeon General.* Rockville: MD: US Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services, National Institutes of Health, National Institute of Mental Health.

van Hasselt M. 2008. Bayesian inference in a sample selection model. Working paper, Department of Economics, The University of Western Ontario.

Wells KB, Burman MA. 1991. Caring for depression in America: lessons learned from early findings of the Medical Outcomes Study. *Psychiatr Med* 9:503-519.

Wynand P, Praag BV. 1981. The demand for deductibles in private health insurance. *Journal of Econometrics* 17:229-252.

Zellner A. 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: P. Joel and A. Zellner (eds.) Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti. Amsterdam: North Holland.